**Clustering on TripAdvisor reviews Restaurant Dataset**

**Group Members:**

|  |  |  |
| --- | --- | --- |
| **First name** | **Last Name** | **Student ID** |
| ***Siddharth*** | ***Yadav*** | ***C0907985*** |
| ***Aniketh*** | ***Vaglapuram*** | ***C0903304*** |
| ***Dev*** | ***Parmar*** | ***C0908036*** |
| ***Guruprasad*** | ***Tandlekar*** | ***C0902813*** |

**Submission date: *March 17, 2024***

**Data:** [ADVPY (DATA)](https://mylambton-my.sharepoint.com/:f:/r/personal/c0908036_mylambton_ca/Documents/ADVPY%20(DATA)?csf=1&web=1&e=4Cw1Un)

Contents

[1. Abstract 4](#_Toc161562630)

[2. Introduction 4](#_Toc161562632)

[3. Methodology 5](#_Toc161562633)

[4. Data Collection: 5](#_Toc161562634)

[5. Data Extraction: 6](#_Toc161562635)

[6. Data Validation & Cleansing: 6](#_Toc161562636)

[7. Data Visualization: 7](#_Toc161562637)

[8. Pandas Profiling: 9](#_Toc161562638)

[9. Text Pre-Processing 9](#_Toc161562639)

[10. Model Building: 11](#_Toc161562640)

[11. Key Insights: 21](#_Toc161562641)

[12. Limitations: 21](#_Toc161562642)

[13. Future Work: 21](#_Toc161562643)

[14. Conclusion: 22](#_Toc161562644)

[15. References: 22](#_Toc161562645)

# **Abstract**

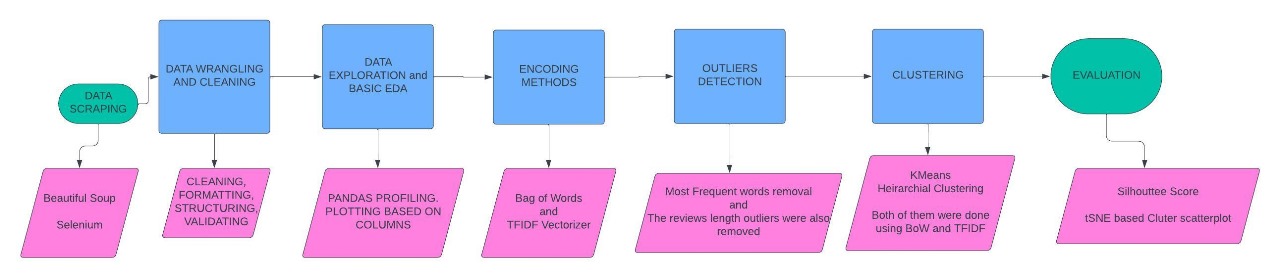
# TripAdvisor.ca relies on user-submitted reviews and ratings for hotels, restaurants, and attractions worldwide. The platform offers search and booking options for accommodations, flights, restaurants, and activities, often featuring exclusive deals. Users access travel guides, itineraries, and destination-specific recommendations for informed trip planning.First, we have scrapped the Data and then performed all the necessary Data cleaning and Feature engineering steps to create some new predictors and properly formatting the scrapped data. Then we have transformed the reviews using various techniques like lemmatization (Converting the Word back to its Root form), Tokenization (Creating tokens of each word from the review) and Text Cleansing (removing stop words, punctuation, Digits, converting all the words to lower case etc.). we also have used multiple visualizations (word clouds, basic bar plot to check distribution of the data, and Panda’s profiling). While exploring the data (EDA) we used techniques like outlier detection and removal to further clean data.

# **Introduction**

In recent years, online platforms such as TripAdvisor have become crucial resources for travelers seeking information about eating experiences. With the growth of user-generated ratings, it can seem difficult to sort through the numerous restaurant selections. To overcome this issue, Natural Language Processing (NLP) is a potential approach. This work uses Text Clustering approaches to, to extract important insights on customer feelings, preferences, and trends. By automating textual data analysis, we can find subtle opinions, allowing diners to make informed selections and restaurants to improve their products to fulfill customer expectations better.

# **Methodology**

The clustering analysis of various customers' evaluations on Tripadvisor is the foundation of this research. The hierarchical and K-means clustering techniques are used. A detailed description of the procedure and how it is carried out is provided below.



*Figure 1.Workflow*

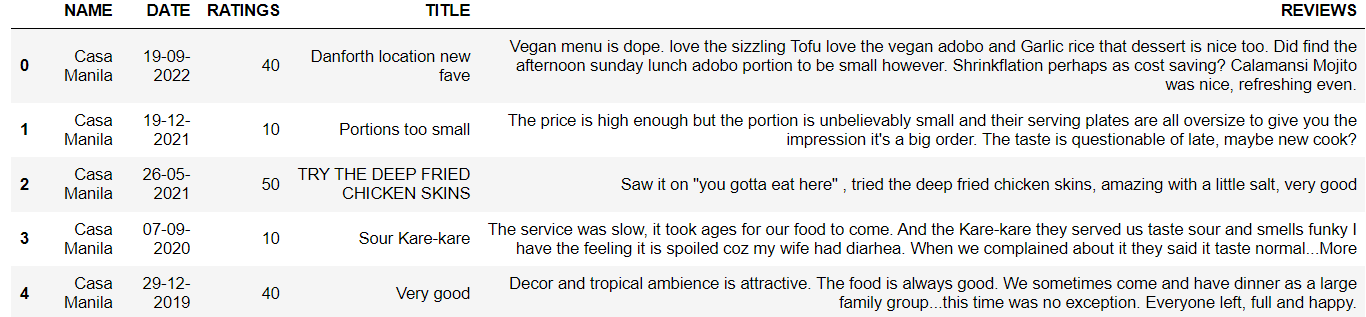
# **Data Collection:**

Based on a restaurant's ranking on the TripAdvisor website, we scraped the data for that restaurant. The dataset comprises six hundred and seventy-six user reviews—name, date, ratings, title, and reviews. This dataset comprises 676 entries and five columns of numerical and categorical data, which was obtained from the TripAdvisor website. The following is a summary of the attribute:

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| Name | obj | Name of the restaurant |
| Date | obj | Date of the review |
| Ratings | Int | Overall rating given by the reviewer (e.g., 5, 3). |
| Title | obj | Title given by the reviewer |
| Reviews | obj | Description of their review |

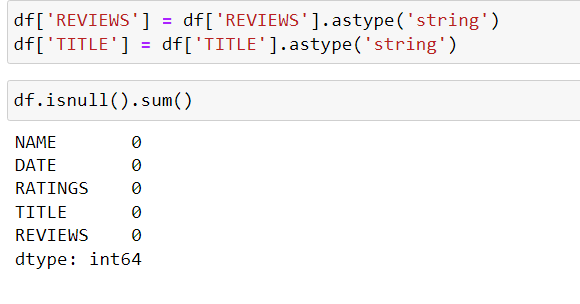
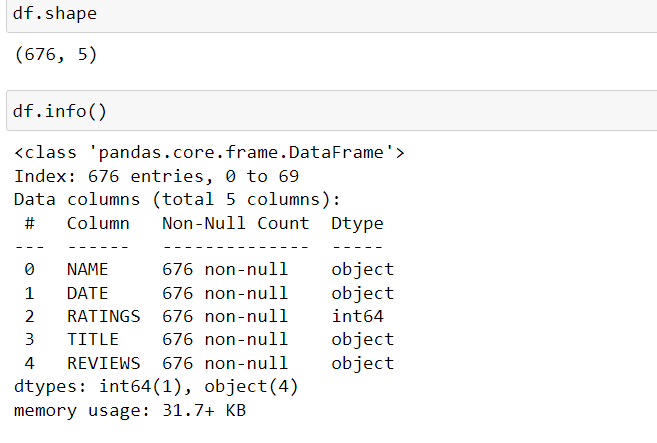
# **Data Extraction:**

Important libraries like SciPy, NLTK, and Matplotlib were loaded for fundamental analysis and data extraction. Matplotlib makes data visualization possible, SciPy gives a range of tools for scientific computing, and NLTK provides natural language processing skills essential for text analysis. And other libraries were there. To enable effective modification and study, the raw data was put into a Pandas Data Frame named "df."



*Figure 1. Trip-Advisor Dataset*

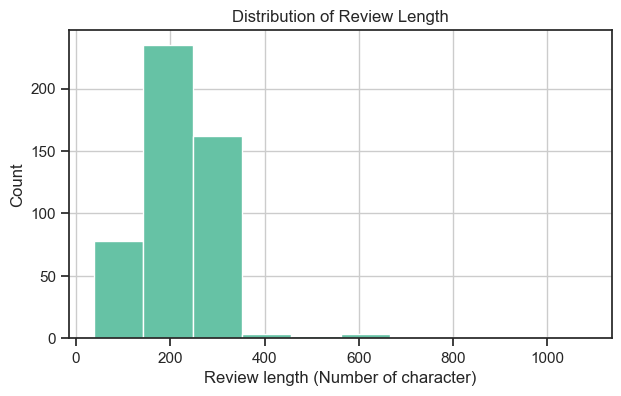
# **Data Cleansing:**

We implemented several data-cleaning techniques to guarantee the accuracy of our dataset. To begin with, redundant data was found and eliminated from the "Reviews" and "Title" columns. The 'Reviews' column data type was also changed to a string from an object to enable uniform data handling and processing. Our dataset was confirmed to be complete by doing a comprehensive check for null values in every column, which showed no missing data.

*Figure2. Datatypes of Columns and Checking Null Values*

# **Data Visualization:**

This bar chart presents the frequency distribution of review lengths. This Clearly shows that reviews are most commonly around 200 characters long, with a notable decline in frequency as the length increases. No reviews exceed 600 characters, highlighting a preference for concise feedback.



*Figure 3. Distribution of rating and review length*

The graph exhibits a consistent upward trend in the counts of all rating categories over the years. This suggests growth or increased activity related to these ratings. Till 2019 All rating categories experience a sudden surge in counts. But after 2019 there was downturn till 2024 which gives only 1 category Review. Investigating the reasons behind this spike could provide valuable insights.

A graph of different colored columns

Description automatically generated

*Figure 4. Trend of Ratings Over Time.*

The graph exhibits a consistent upward trend in the counts of all rating categories for the restaurants. This suggests growth or increased activity related to these ratings. However, McDonald's has fewer restaurants on the list compared to all the other restaurants on the list. Investigating the reasons behind these changes could provide valuable insights.

A graph of different colored lines

Description automatically generated

*Figure 4. Trend of Ratings on Restaurants.*

# **Pandas Profiling:**

We have added all our data into pandas profiling. It provides a quick and efficient way to understand the structure and characteristics of our dataset, including data types, missing values, distribution of values, correlation between variables, and more. This generated report making it easy for us to analyses to gain insights into the data quickly. Before moving further with any analysis or modelling, Pandas Profiling is very helpful in the early phases of data exploration and preprocessing, assisting us in seeing possible problems and trends in the data.

A screenshot of a computer

Description automatically generated

*Figure 5.Pandas Profiling.*

# **Text Pre-Processing**

When natural language processing (NLP) is used for preprocessing activities, text cleaning is essential in improving the quality and efficiency of the resulting analysis by refining the raw textual input. This section describes the approaches and strategies used in the text-cleaning process to improve data quality and enable precise linguistic analysis. Some of the Text cleaning Techniques we used are:

* **Removal of Stop Words:** Stop words, such as "is," "the," and "and" were removed from the text as they do not convey meaningful information for clustering analysis.
* **Punctuation and Special Characters Removal:** To maintain uniformity in text processing, special letters and punctuation were removed from the text.
* **Removing Digits:** Numeric digits were removed from the text as they typically do not contribute to the sentiment or topics of the reviews.
* **Lemmatization:** Lemmatization was used to simplify the vocabulary and increase the precision of clustering by breaking down terms in the text into their base or root form.
* **Word Cloud:** To provide insight into the most frequently occurring keywords, a word cloud was created to depict the frequency of words in the text data graphically.
* **Emoji Removal:** To remove the reviews containing emojis.
* **Removing Meaningless Words:** This function was created to correct spelling using spellchecker from Text Blob.

Word Cloud

A word cloud was used as a visualization tool in our assignment examining review analysis. By displaying the most frequently occurring terms or themes in the evaluations graphically, this word cloud was used to summarize their main points. With this method, we could quickly show the themes and opinions in all assessments, making it easier to analyze and understand the deeper data.

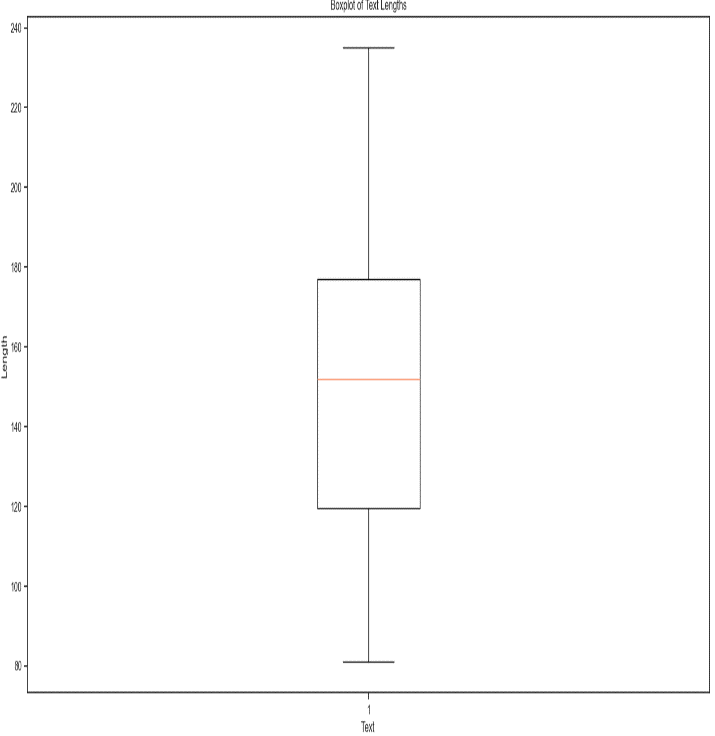
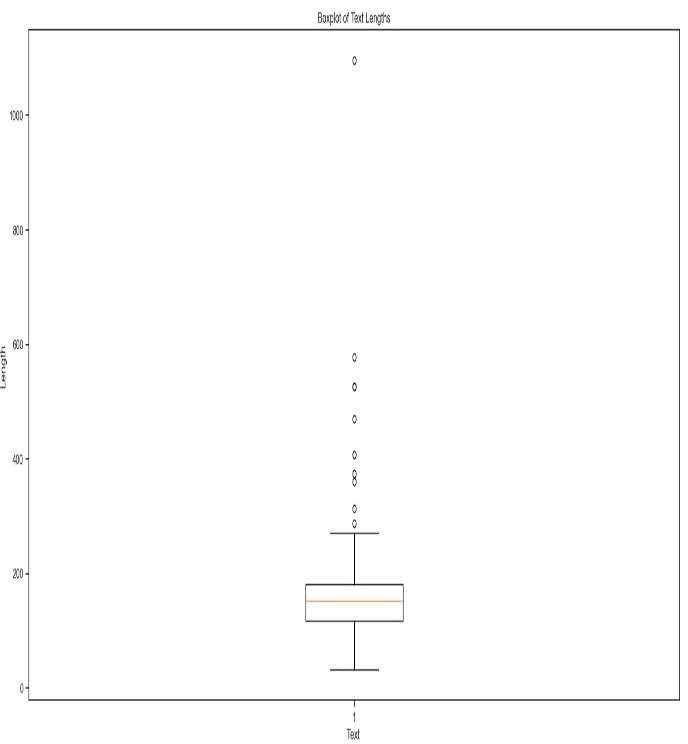


*Figure 6.Word Cloud.*

# **Outlier Detection and Removal**

In This length of reviews for Outlier Detection reviews. Reviews that were longer than 300 words were considered as outlier. For Outlier removal We used Three method First Trimming (Trimming involves removing a certain proportion of extreme values from both ends of the dataset.), Quantile-based Flooring and Capping(This technique sets thresholds based on percentiles (quantiles) of the data and caps or floors values beyond these thresholds.), Lastly Log Transformation (Log transformation involves taking the logarithm of each observation in the dataset.).

**BOXPLOT:**



*Figure 6.Before and After Removal Outlier Removal.*

# **Clustering Method:**

This part involves applying two widely used clustering techniques to the preprocessed restaurant review dataset: K-Means and Hierarchical Clustering. By combining comparable evaluations regarding linguistic content, these algorithms sought to reveal underlying patterns and structures in the data.

1. **K-Means Clustering:**

While K-Means clustering is straightforward and computationally efficient, it depends on predetermined cluster centroids and is sensitive to initializations, frequently leading to less-than-ideal results. The elbow approach was used to choose the number of clusters.

1. **K-Means Clustering with Bag of Words (BoW):**

We employed the elbow approach after starting K-Means clustering with a range of cluster numbers from 2 to 15 to find the ideal number of clusters.

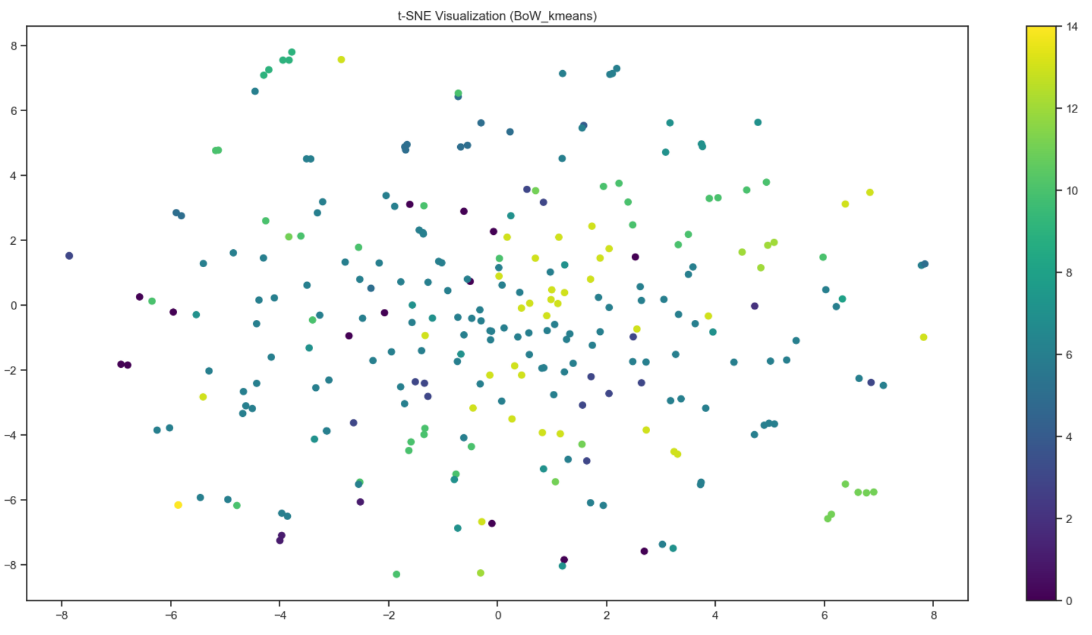
A graph with a line

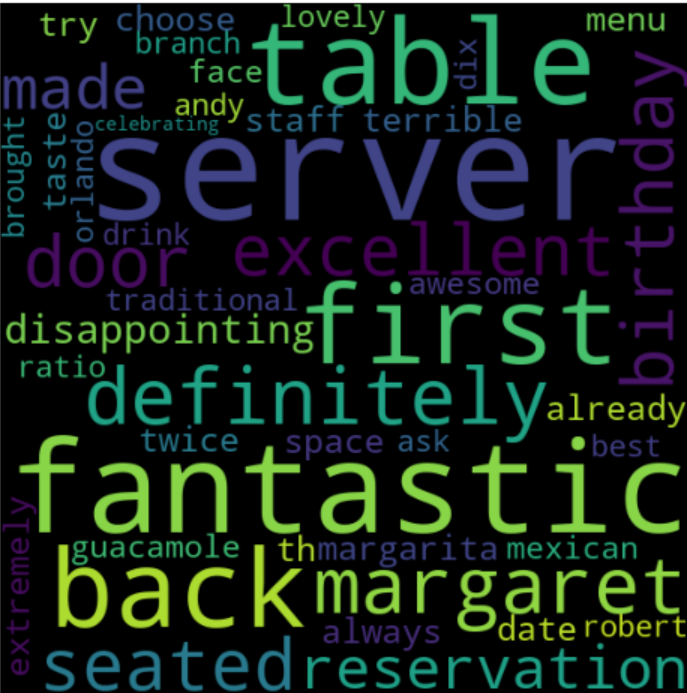
Description automatically generated

*Figure 7.Elbow Method*

**Results:**

We have T Distributed Stochastic Neighbour Embedding (t-SNE) to decrease the dimensionality of the data to two dimensions so that the clustering results could be visually inspected. The t-SNE visualization, however, showed overlapping clusters despite the clustering analysis, suggesting possible limits in cluster distinctiveness and the existence of mixed memberships inside clusters.

****

*Figure 8.T-distributed stochastic neighbor embedding.*

*Figure 8. Word cloud for review 1 in a cluster of BOW.*

**Evaluation Metrics:**

Using the silhouette score, which measures how cohesive and distinct groups are from one another, we evaluated the quality of clustering. A higher score denotes better-defined clusters; the silhouette score runs from -1 to 1. The silhouette score of 0.021014077732904875 for BoW-based K-Means clustering indicates minimal group separation.

**

1. **K-Means Clustering with TF-IDF Representation:**

Similar to BoW representation, K-Means clustering was applied to the TF-IDF representation of the dataset. We employed the elbow approach after starting K-Means clustering with a range of cluster numbers from 2 to 15 to find the ideal number of clusters.

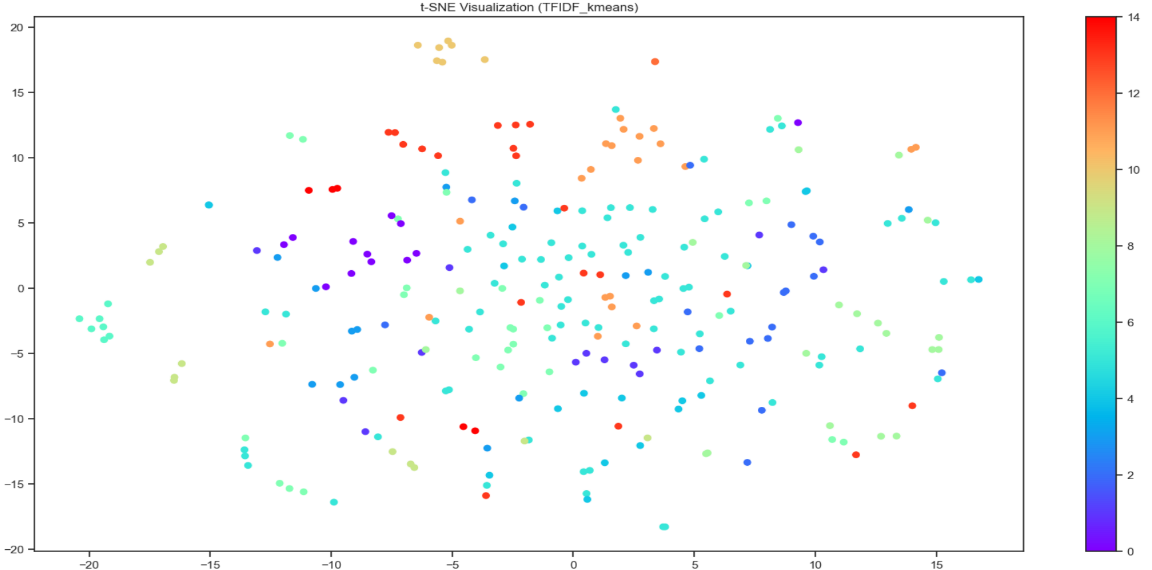
A graph with a line

Description automatically generated

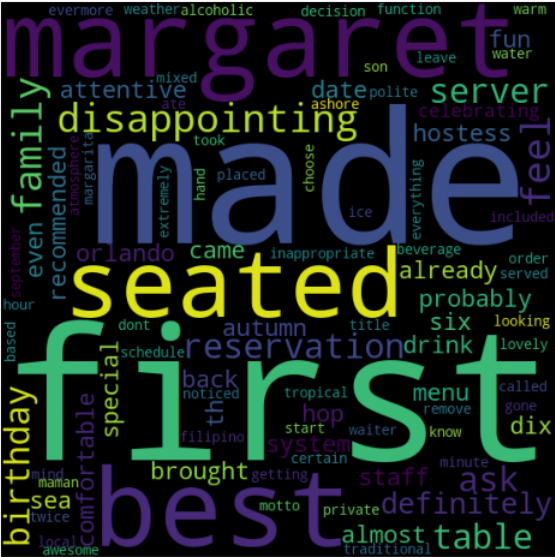
*Figure 9.Elbow Method in TFIDF*

**Results:**

Similarly, we used t-SNE to show the clustering findings and investigate the two-dimensional data point distribution. On the other hand, overlapping clusters were also seen in the t-SNE visualization for TF-IDF-based K-Means clustering, indicating difficulties in attaining distinct cluster separation.

****

*Figure 10.T-distributed stochastic neighbor embedding*

****

*Figure 11. Word cloud for review 1 in a cluster of TFIDF.*

**Evaluation Metrics:**

For TF-IDF-based K-Means clustering, we calculated the silhouette score in order to evaluate the quality of the generated clusters. The silhouette score showed somewhat greater separation across clusters, improving to 0.04127434275248862 when compared to BoW representation.

****

1. **Hierarchical Clustering:**

Hierarchical Clustering offers a more flexible approach by representing the data in a hierarchical tree-like structure, allowing for the exploration of different levels of granularity in clustering. However, hierarchical clustering can be computationally intensive, particularly for large datasets, and the choice of linkage method and distance metric can significantly impact the resulting clusters.

1. **Hierarchical with Bag of Words (BoW):**

We used the Ward linkage approach, which reduces variance while merging clusters, for Hierarchical Clustering using the Bag of Words (BoW) representation. This approach, which seeks to create compact and spherical clusters, is especially well-suited for high-dimensional data, such as text data. Using dendrograms, which show the merging process step-by-step, we could illustrate the hierarchical relationships between data points.

A colorful lines in different colors

Description automatically generated

*Figure 12**A Hierarchical Clustering Dendrogram*

**Cluster Interpretation:**

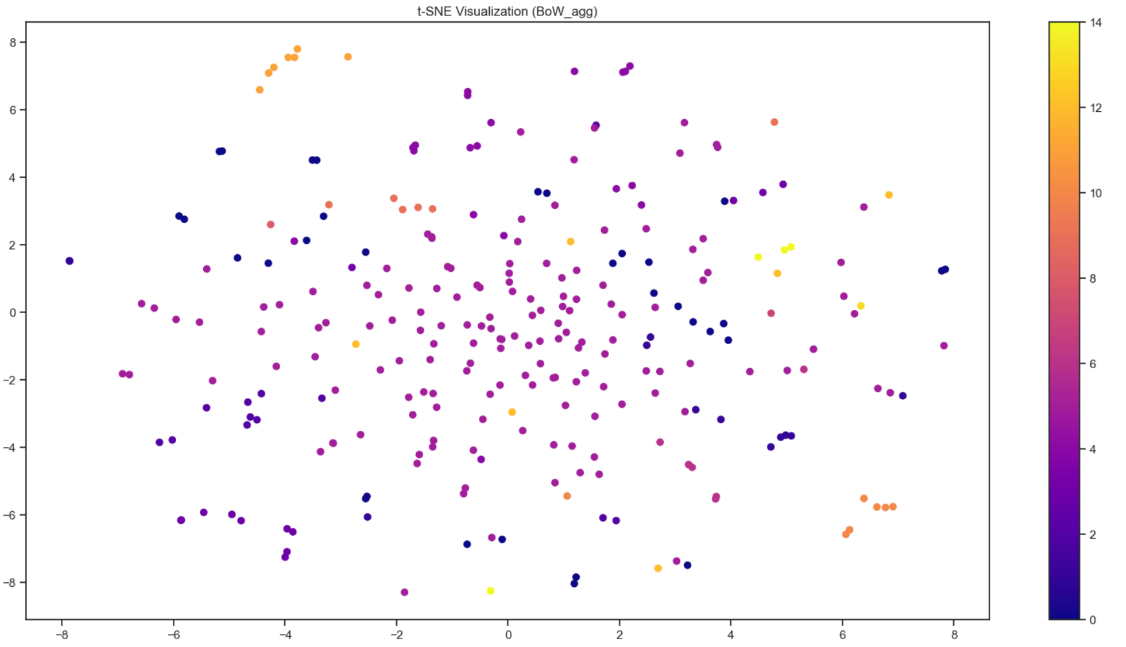
We identified the appropriate number of groups and then extracted them using hierarchical clustering. The distribution of reviews within each of the 15 clusters we found throughout our study was examined. We saw that the sizes and features of the clusters varied, with some having more reviews than others.

A number grid with numbers

Description automatically generated with medium confidence

*Figure 13**Extracted Clusters*

**Results:**

Using t-distributed Stochastic Neighbor Embedding (t-SNE), we could display the clustering findings and minimize the number of dimensions in the data to two. However, like with K-Means clustering, the t-SNE visualization showed overlapping clusters, indicating possible difficulties obtaining clear cluster separation.

*Figure 14.T-distributed stochastic neighbor embedding*

****

*Figure 15. Word cloud for review 15 in cluster of BOW.*

**Evaluation Metrics:**

The cohesiveness and separation of clusters are measured by the silhouette score, which we computed to assess the quality of clustering. It was discovered that the silhouette score for BoW-based Hierarchical Clustering was 0.06069685008829494, which indicates a modest degree of cluster separation.

**

1. **Hierarchical with TF-IDF Representation:**

Likewise, we used the Ward linkage approach to apply Hierarchical Clustering to the dataset's TF-IDF representation. To comprehend the merging process and cluster structures, we used dendrograms to show the hierarchical relationships between data points.

A colorful lines in different colors

Description automatically generated

*Figure 16. A Hierarchical Clustering Dendrogram*

**Cluster Interpretation:**

We found 15 groupings using hierarchical clustering and looked at how reviews were distributed across them. We saw that the clusters varied in size and composition, indicative of the range of subjects and opinions included in the evaluations.

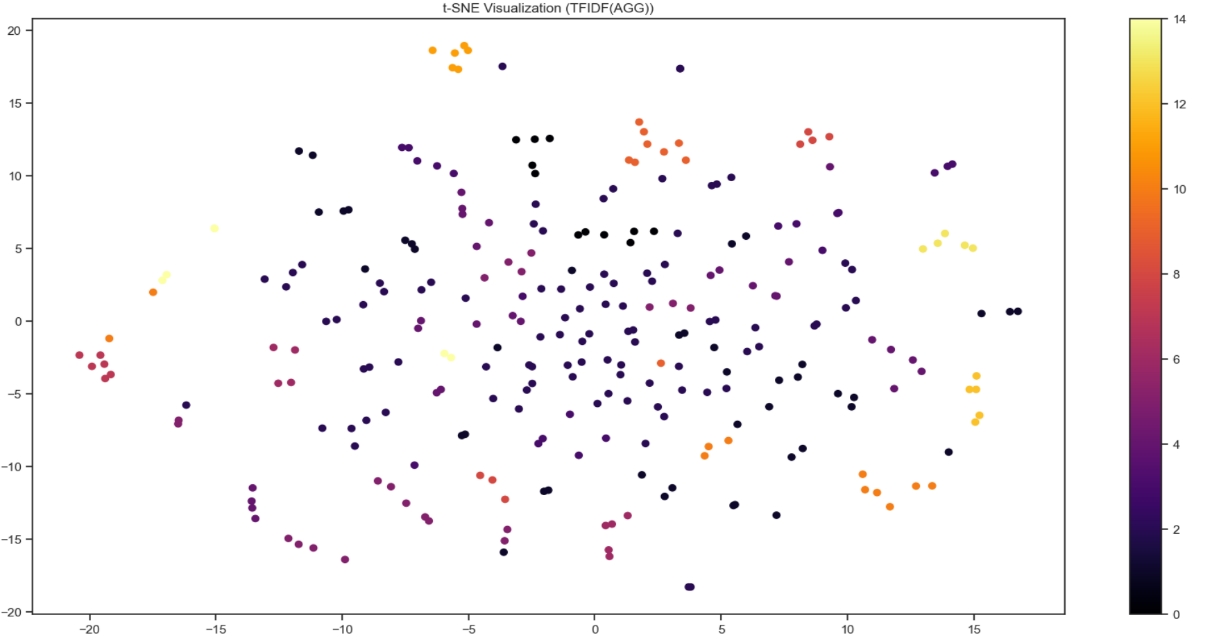
*A number grid with numbers and symbols

Description automatically generated with medium confidence*

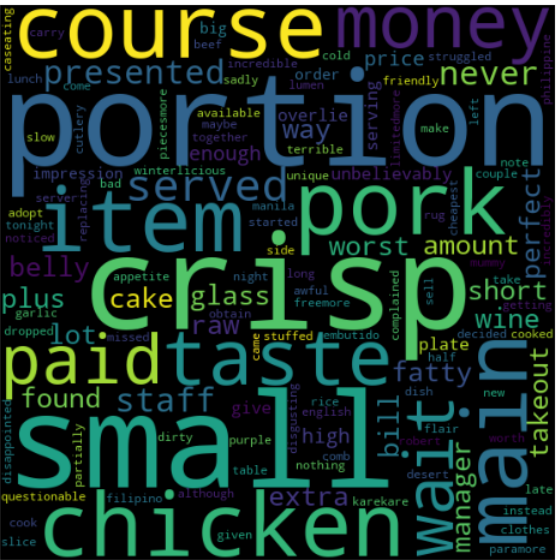
*Figure 17.Extracted Clusters*

**Results:**

We visualized the clustering findings in two dimensions using t-SNE to investigate the distribution of data points within clusters. On the other hand, the t-SNE visualization showed overlapping clusters, which suggested possible difficulties in attaining distinct cluster separation, much as BoW-based Hierarchical Clustering.

**

*Figure 18.T-distributed stochastic neighbor embedding*

****

*Figure 19. Word cloud for review 1 in cluster of TFIDF.*

**Evaluation Metrics:**

We computed the silhouette score for TF-IDF-based Hierarchical Clustering to evaluate the clustering quality. Comparing the silhouette score to the BoW representation, it increased somewhat to 0.05567670374544527, showing better cluster separation.

****

# **Key Insights:**

Using the Bag of Words (BoW) and TF-IDF representations, we applied the K-Means and Hierarchical Clustering methods to find unique clusters in the dataset. These clusters indicated food quality, service standards, ambiance, and cost were just a few of the varied consumer experience elements. Our research provided detailed knowledge of the variables influencing restaurant evaluations by highlighting underlying patterns and trends in consumer opinions and preferences.

# **Limitations:**

It is crucial to recognize some limitations even if our clustering approach has yielded insightful results. Accurately classifying reviews might be difficult due to overlapping themes and ambiguous cluster borders. Moreover, pretreatment processes determine how successful clustering approaches are, emphasizing the necessity of reliable text cleaning and feature extraction techniques.

# **Future Work:**

These results open several directions for further investigation. Tests using more complex text embedding methods, such as Word2Vec or BERT, improve the clustering accuracy and the representation of textual data. The shortcomings of individual algorithms may also be addressed, and more refined results may be obtained by investigating ensemble clustering techniques and using domain knowledge. Sentiment research combined with clustering may offer a more comprehensive understanding of consumer attitudes and support focused company strategy formulation.

# **Conclusion:**

We have learned a great deal about the preferences and opinions of customers from our study on clustering approaches applied to restaurant evaluations. Although obstacles and prospects exist for enhancement, the results provide practical perspectives that might guide strategic choices in the restaurant sector. We can uncover more insights and promote ongoing improvements in customer satisfaction and corporate success by carrying out new approach research and methodological refinement.

# **References:**

1. Han, J., & Kambar, M. (2006). Data Mining: Concepts and Techniques. Morgan Kaufan.
2. Pedregose, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 13, 2826-2830.
3. Yelp Dataset Challenge. (n.d.). Retrieved from <https://www.yelp.com/dataset>.
4. Jain, A. K., Murthy, M. N., & Flynn, P. J. (1998). Data clustering: A review. ACM Computing Surveys (CSUR), 31(3), 264-323.